Der Korrelationskoeffizient liegt zwischen –1 und +1

HANS RIEDWYL, BERN

Zusammenfassung: *Mit dieser Arbeit wird auf einfache Weise gezeigt, dass der Korrelationskoeffizient immer zwischen –1 und +1 ist.*

1 Einleitung

In vielen Lehrbüchern und Skripten der Statistik wird unter den Eigenschaften des Korrelationskoeffizienten vermerkt, dass dieser zwischen –1 und +1 liegen muss. Häufig fehlt eine Herleitung oder es wird auf die Cauchy-Schwarz'sche Ungleichung verwiesen, die für Schüler und Studenten in der Einführungsphase des Studiums wenig bekannt ist. Der folgende Beweis vollzieht den Beweis der Cauchy-Schwarz'schen Ungleichung in der hier benötigten Form. Viele Texte beginnen zudem mit deskriptiver Statistik und möchten diese Eigenschaft ohne den Begriff der Zufallsvariable vorauszusetzen beweisen.

Im Abschnitt 2 wird gezeigt, dass der empirische Korrelationskoeffizient zwischen –1 und +1 liegt und dass er exakt –1 oder +1 ist, wenn die Wertepaare auf einer Geraden liegen. Im Abschnitt 3 wird dasselbe für den theoretischen Korrelationskoeffizient demonstriert.

2 Empirisch

Gegeben seien n Wertepaare (x_i, y_i), i=1,2,...,n. Der empirische Korrelationskoeffzient ist definiert als

$$r_{xy} = \frac{c_{xy}}{s_x s_y} = \frac{c_{xy}}{\sqrt{c_{xx} c_{yy}}} \qquad (1)$$

wobei c_{xy} die Kovarianz

$$c_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}),$$

$s_x^2 = c_{xx}$ und $s_y^2 = c_{yy}$ die Varianz von x resp. y sind. Wir schließen dabei aus, dass eine oder beide Varianzen Null sind.

Der Korrelationskoeffizient (1) lässt sich auch darstellen als

$$r_{xy} = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \quad . \qquad (2)$$

In den runden Klammern stehen die sog. Standardwerte bezüglich x resp. y.

Für n > 1 ist die folgende Ungleichung stets gültig:

$$\frac{1}{n-1}\sum[\frac{x_i-\bar{x}}{s_x}+\frac{y_i-\bar{y}}{s_y}]^2 \geq 0 . \qquad (3)$$

Quadrieren wir das Binom in der eckigen Klammer, so folgt daraus mit

$$\frac{1}{n-1}\sum[\frac{x_i-\bar{x}}{s_x}]^2 = \frac{1}{n-1}\sum[\frac{y_i-\bar{y}}{s_y}]^2 = 1$$

die Ungleichung

$$1+1+2r_{xy} \geq 0$$

und daraus unmittelbar

$$-1 \leq r_{xy} .$$

Ersetzen wir in (3) in der eckigen Klammer das Pluszeichen durch ein Minuszeichen, so folgt daraus sofort, dass

$$r_{xy} \leq 1$$

sein muss und damit

$$-1 \leq r_{xy} \leq 1 \qquad (4)$$

Der empirische Korrelationskoeffizient ist genau dann +1, wenn die Standardwerte in x und y alle identisch sind, d.h. wenn für alle i gilt:

$$\frac{x_i-\bar{x}}{s_x} = \frac{y_i-\bar{y}}{s_y} .$$

Das ist nur der Fall, wenn alle Wertepaare (x_i, y_i) auf einer Geraden mit positiver Steigung liegen. Analog kann gezeigt werden, dass der empirische Korrelationskoeffizient genau dann –1 ist, wenn die Summe der Standardwerte von x und y für alle Wertepaare Null ist, d.h. die Punkte liegen auf einer Geraden mit negativer Steigung.

3 Theoretisch

Gegeben zwei Zufallsvariablen X und Y mit den Erwartungswerten $E[x]= \mu_x$ resp. $E[y]= \mu_y$, den Varianzen $Var[x]=\sigma_x^2$ resp. $Var[y]=\sigma_y^2$ und dem theoretischen Korrelationskoeffizienten

$$\rho_{xy} = E[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}] . \qquad (5)$$

Wir schließen dabei aus, dass eine der beiden Zufallsvariablen mit Wahrscheinlichkeit 1 konstant ist.

Da der Erwartungswert positiver Größen stets positiv sein muss, ist

$$E[(\frac{x-\mu_x}{\sigma_x} \pm \rho_{xy}\frac{y-\mu_y}{\sigma_y})^2] \geq 0 . \qquad (6)$$

Da die Erwartungswerte

$$E[(\frac{x-\mu_x}{\sigma_x})^2] = E[(\frac{y-\mu_y}{\sigma_y})^2] = 1$$

sind, folgt $1+1\pm 2\rho_{xy} \geq 0$ oder

$$-1 \leq \rho_{xy} \leq +1 . \qquad (7)$$

Möchte man zusätzlich auch zeigen, dass die gemeinsame Verteilung von X und Y in den Extremlagen $\rho_{xy} = \pm 1$ auf einer Geraden konzentriert ist, kann nach van der Waerden [1] eine neue Zufallsvariable

$$Z = Y - \beta X \qquad (8)$$

eingeführt werden, für die gilt:

$$\sigma_z^2 = \sigma_y^2 + \beta^2\sigma_x^2 - 2\beta\rho_{xy}\sigma_x\sigma_y . \qquad (9)$$

Durch elementares Differenzieren und Nullsetzen findet man, dass die quadratische Funktion in β ihr Minimum an der Stelle

$$\beta = \rho_{xy}\frac{\sigma_y}{\sigma_x} \qquad (10)$$

hat. Eingesetzt in (9) wird das Minimum

$$\sigma_z^2 = \sigma_y^2(1-\rho_{xy}^2) . \qquad (11)$$

Da die Varianz von Z nicht negativ sein kann, folgt unmittelbar wiederum die Ungleichung (7).

Für $\rho_{xy} = \pm 1$ ist $\sigma_z^2 = 0$, was nur möglich ist, wenn Y mit Wahrscheinlichkeit 1 auf einer Geraden mit Steigung β ist.

Setzt man an Stelle von (8) die Gleichungen

$$z_i = y_i - \beta x_i \quad , i = 1,2,...,n \qquad (12)$$

und folgt dem gleichen Weg mit empirischen Wertepaaren, so erhält man auch die Resultate des Abschnittes 2.

Nach den Beziehungen (8) bis (12) vorzugehen, empfiehlt sich besonders dann, wenn das Modell der linearen Regression bereits eingeführt wurde. Denn β ist dann gleich dem Regressionskoeffizienten und (10) zeigt, dass β und ρ_{xy} immer dasselbe Vorzeichen haben. Das Quadrat von ρ_{xy} ist gleich dem Bestimmtheitsmaß und misst nach (11) den Anteil der Varianz von y , der von der Steigung der Geraden herrührt.

Literatur

van der Waerden, B.L. (1957). Mathematische Statistik. Springer-Verlag, Berlin.

Adressen des Autors
Prof. Dr. Hans Riedwyl
Universität Bern
Inst. f. Math. Statistik und Versicherungslehre
Sidlerstrasse 5
CH-3012 Bern
hans.riedwyl@stat.unibe.ch

und

Consult AG Bern
Statistische Beratung
Kirchstrasse 40
CH-3097 Liebefeld
hans.riedwyl@consultag.ch